DOCUMENT RESUME

ED 427 522 FL 025 694

AUTHOR Pino, Barbara Gonzalez
TITLE Prochievement Testing of Speaking: Matching Instructor Expectations, Learner Proficiency Level, and Task Type.
ISSN ISSN-0898-8471
PUB DATE 1998-00-00
NOTE 17p.; For the complete volume of working papers, see FL 025 687.
PUB TYPE Journal Articles (080) -- Reports - Research (143)
JOURNAL CIT Texas Papers in Foreign Language Education; v3 n3 p119-33 Fall 1998
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Instruction; *Evaluation Criteria; Higher Education; Interrater Reliability; *Language Proficiency; Language Research; Language Teachers; *Language Tests; Oral Language; Second Language Learning; *Second Languages; *Spanish; Speech Skills; Surveys; Teacher Attitudes; *Teacher Expectations of Students; Test Items

ABSTRACT

Previous literature on classroom testing of second language speech skills provides several models of both task types and rubrics for rating, and suggestions regarding procedures for testing speaking with large numbers of learners. However, there is no clear, widely disseminated consensus in the profession on the appropriate paradigm to guide the testing and rating of learner performance in a new language, either from second language acquisition research or from the best practices of successful teachers. While there is similarity of descriptors from one rubric to another in professional publications, these statements are at best subjective. Thus, the rating of learners' performance rests heavily on individual instructors' interpretations of those descriptors. An initial investigation of instructor assumptions was conducted regarding student performance on speaking tests in one program and identified several areas of discrepancy in instructor testing and rating practice. It is argued that faculty as a group must delineate more clearly their specific expectations by level for a number of rated features. The concerns identified in this study coincide with those discussed recently in the literature, suggesting that other programs might benefit from similar self-analysis. The instructor questionnaire is appended. Contains 17 references. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Prochievement Testing of Speaking: Matching Instructor Expectations, Learner Proficiency Level, and Task Type

BARBARA GONZALEZ PINO, The University of Texas at San Antonio

ED 427 522

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Prochievement Testing of Speaking: Matching Instructor Expectations, Learner Proficiency Level, and Task Type

BARBARA GONZALEZ PINO, The University of Texas at San Antonio

Earlier literature on classroom testing of speaking provides several models of both task types and rubrics for rating and suggestions regarding procedures for testing speaking with large numbers of learners. There is no clear, widely disseminated consensus in the profession, however, on the appropriate paradigm to guide the testing and rating of learner performance in a new language, neither from second language acquisition research nor from the best practices of successful teachers. While there is similarity of descriptors from one rubric to another in professional publications, these statements are at best somewhat subjective. Thus, the rating of learners' performance rests heavily on individual instructors' interpretations of those descriptors. The author conducted an initial investigation of instructor assumptions regarding student performance on speaking tests in her own program and identified several discrepant areas of instructor testing and rating practice. Further, faculty as a group will need to delineate more their specific expectations by level for a number of the rated features. The concerns identified coincided with those discussed recently in the literature, which suggests that other programs may also benefit from similar self-analysis.

INTRODUCTION

The language educator who is familiar with both the American Council on Teaching Foreign Languages (ACTFL) Proficiency Guidelines and the proficiency levels typically achieved by first- and second-year university learners of foreign languages may well have questions about some of the ways in which speaking is tested and rated in classes in some of those university programs. If learners are in the Novice Mid to Intermediate Mid range, are the testing tasks they are given always level-appropriate and should they be? Are the descriptors of the rating scales used by the instructors always appropriate and clear? Do all the instructors interpret general or ambiguous descriptors in the same way? Is the system of testing and rating that is in place implemented consistently by the instructors in a program and should it be? Do the tests match well with what is covered in classes? Do the instructors' expectations while rating reflect a solid understanding of what learners can do at their proficiency level? Do their expectations reflect an understanding of the restructuring of knowledge and performance that may occur in learners as they move to the Intermediate proficiency level? How do instructor-raters handle beyond-level tasks? We may not always know the answers to these questions as they pertain to our own programs, much less on a larger scale.

Concern about these questions and about the issue of fairness to students led the author to conduct a study in the lower-level program of her own institution to determine the procedures and criteria her instructors were actually using in testing and rating speaking and the extent to which these criteria coincided with program goals and current best practice. Clearly, as in most programs, instructors were testing tasks that were above the students' proficiency level, and their rating scales did not really distinguish level-appropriate tasks from above-level tasks. In addition, the degree of possible variations in implementation and interpretation of process, procedure, and criteria in testing and rating was unknown. Therefore, the author proposed to investigate these areas and use the results to initiate discussion among instructors in the program and in a broader professional setting.

THE LITERATURE

Throughout the recent literature on the testing of speaking, many concerns are raised about testing and rating procedures. Much of this literature, however, focuses on proficiency testing rather than on prochievement testing (Clark, 1989), a term that refers to the kind of proficiency-oriented achievement testing we do in foreign language classes on a regular basis, perhaps several times a semester. Nevertheless, in both the proficiency literature and the prochievement literature, aspects of testing and rating appropriately or inappropriately are discussed at length, and some of the features discussed in proficiency studies may have relevance for prochievement testing as well. In general, the profession defines achievement tests as those limited to a particular body of material just covered in class(es) and proficiency instruments as those testing the total range of skills and contexts a learner may be able to handle—regardless of where and when they may have been learned—and testing them through actual interaction in realistic situations. Prochievement tests are a combination of the preceding two types, testing students' ability to perform in only the contexts and situations that have been practiced in class.

Formats

A common concern is whether particular tasks or formats are best suited to certain proficiency levels or particular teaching and testing circumstances. According to Fulcher (1996), there is little evidence to suggest that any particular task format is more suited to one proficiency level than another. Indeed, most of the common formats can be used with learners at different points in their studies. With appropriate expectations, the picture, the topic, the interview, the multi-skilled or integrated task, and even the more demanding roleplay can all be adapted according to the level of the students. The crucial element in using the formats well is their content, which should comprise functions, topics or life situations, and grammatical features appropriate for the particular students and the material covered in their classes (Gonzalez Pino, 1989).

Instructor-Rater Expectations

The aspects of the topic of testing speaking that are most thoroughly covered in the literature are those of rating and the underlying instructor expectations that are such an important part of rating. Effective ways to rate have been studied for decades and in an extensive variety of formats and weighting schemes (Hart Gonzalez, 1994). Thompson (1996) found that rating may vary according to whether the test is taped or not and noted that a rater who is listening to a tape is not as likely to be distracted by the human qualities present in a live interview and is more likely to pay greater attention to form. Richards and Chambers (1996) studied a number of instructor-related variables in rating and reported that many of them have some effect on the rating process. They stated that training on how to rate improves consistency and that linguistic background counts, because native speakers rate more stringently. According to their study, the type of school in which teachers work matters; teachers in more elite or selective schools rate more stringently. One type of teacher experience significant in their findings is experience with learners at the level being rated. Length of overall teaching experience does not matter, however.

Richards and Chambers (1996) examined three types of rating scales in their studies: a norm-referenced categorical scale (one with weighted criteria and numeric scales for each but with no descriptors for the criteria), a criterion-referenced categorical scale (one with a set of criteria, each with a hierarchy of descriptors and numeric values), and a global criterion-referenced scale (one with descriptors and numeric values for each of several general levels of performance). They found that the two more global scales were more reliable, but they explained their finding by suggesting that the descriptors for the criterion-referenced categorical scale were vague and would require much greater specificity in order to function appropriately. Douglas (1994) found that most raters who used scoring rubrics were apparently affected by aspects of performance that were not mentioned in the rubrics. He noted that grammar and rhetorical complexity were particular problem areas for which teacher-raters might employ their own standards or substandards. Richards and Chambers (1996) discovered that pronunciation and grammar caused the greatest rating problems in their study, possibly because these two areas are concrete and yet have no specific detailed standards set out in common for the various levels for all raters to use.

In their 1995 study, Chambers and Richards also found that if the criteria to be used in rating were not described in some detail, teachers varied in their interpretations of the descriptors. Further, they found that teachers may expect strong performance on grammatical elements even if those elements are not appropriate to the task, are not appropriate to the students' level, and would not have been used by native speakers on the same task. Their study specifically compared learners' and native speakers' performances on the same set of tasks in order to compare the grammatical structures that were

used. They determined that teachers may expect forms that not even native speakers employ. They also found that learners who spoke more often received higher ratings, regardless of quality issues. Finally, they determined that these expectations frequently persisted despite differing expectations written into course syllabi, where certain features were cited for recognition and others for both recognition and production.

Thompson (1995) found that a group of proficiency raters tended to develop idiosyncratic testing and rating procedures as compared to other groups. Mullen (1978) recommended that more than one rater rate each test in order to eliminate the effect of rater inconsistency, a practice that has often been followed in proficiency testing since that time, although that procedure would not be practical in classroom testing multiple times per semester. Ross (1987) raised the issue of the appropriate mental construct to undergird norm-referenced scales, suggesting the proficient nonnative would be a better standard than the educated native speaker and highlighting the fact that we may vary in the standard to which we refer. Meredith (1990) suggested further that when rating, teachers must consider whether or not learners have had prior experience in the language; thus, he indicates yet another way in which our mental model and our expectations may vary. Whom do we expect our learners to be like? And do we expect a higher performance level of our false beginners than that indicated for all learners in a particular course?

Levels of Proficiency

Several other concerns in the literature center on the proficiency levels themselves. Stansfield and Kenyon's (1992) study reported that Intermediate and Advanced tasks are more difficult to rate than Novice and Superior and that sublevels of performance in the midranges are more problematic to distinguish from each other. Byrnes (1987) pointed out that Intermediates may make more errors than Novice Lows and Novice Mids, a seeming inconsistency. This indication, however, is related to Young's (1992) indication that there may be an uneven progression in language acquisition, even a "U-shaped" phenomenon in which Intermediate learners may seem to regress because, as they acquire new structures and vocabulary and reformulate their interlanguage, the restructuring destabilizes their performance for a time. The fact that they are creating in the language and relying less on memorized material has a similar effect. Thus, in addition to considering whether our expectations of learners are generally appropriate to their level, we must also consider the extent to which those expectations take into account these additional complexities in second language acquisition and in the rating process.

Textbooks

Finally, we can consider our textbooks as a type of professional literature to be examined and having clear implications regarding proficiency levels. First-year textbooks, so called whether they are used for the first year or a year and a half, invariably cover much of the structure

of the language in question and include functions and content that would be consistent with the Advanced and Superior proficiency levels, despite the fact that no learners (other than native speakers, perhaps) are expected to achieve those levels of proficiency during the first-year (or year-and-a-half) course. Second-year materials also typically include Intermediate through Superior material. The case can certainly be made that we are introducing materials at those levels as a pedagogical strategy to enable learners to begin to develop those particular skills. In each program, however, we still must decide on the appropriate way to evaluate performance on Advanced- and Superior-level material relative to performance on functions, structures, and topics for lower levels of proficiency, both when we design and when we rate our prochievement tests of speaking.

Summary of Literature

Clearly, then, in summary, the literature addresses some of our initial concerns by indicating that many variables affect teachers' rating of learners' speaking. Chief among these variables is the teacher's own set of expectations of students. Since these expectations could apply even in the face of specific statements on syllabi constraining such expectations and despite recommendations contrary to instructor expectations presented during training on how to rate, the concerns seem valid. Since only Richards and Chambers' (1996) and Gonzalez Pino's (1989) studies specifically concern class-related testing, while the others focus on proficiency testing of speaking, however, the investigator undertook to explore further the extent of variation in expectations among instructors who rate their own students' prochievement tests of speaking.

DESCRIPTION OF THE SURVEY

Several researchers have mentioned the need to ask raters to engage in self-assessment and in "think-aloud" protocols. They point out that a comparison of assigned ratings does not always permit an analysis of underlying differences in expectations since two raters can assign the same score for different reasons. Thus, this investigation begins with a self-assessment of testing and rating procedures that, it is hoped, will pinpoint further topics for investigation. The author will analyze consistency among raters and the relationships of responses to the constructs of the ACTFL Proficiency Guidelines and the tenets of second language acquisition

The Sample

Twenty instructors of lower-level language courses at the university level participated in the survey. These individuals teach in a communicatively oriented program in which the policy calls for daily emphasis on speaking skills. They administer and rate speaking tests for their own learners three times each semester. A set of 20 to 30 sample oral test items is provided to the instructors and the learners 2 weeks prior to each test. Each test comprises pictures, topics, interviews, and role-plays related to the chapters in question. The items, which were developed by a subcommittee of instructors for use by the entire group of 20

to 25 for coordinated examinations, cover the dozens of functions and topics included in the text. The text includes functions and topics appropriate to the Advanced and Superior levels reflective of the texts discussed above. The instructors then use the sample items as a repertoire or bank of items as they individually structure the way in which they will administer the test. The instructors determine whether they will test one-on-one or in learner pairs, in class or in their offices, whether they will use two or more formats on a given test (two are the departmental minimum), and whether they will tape record the test or not.

The instructor-raters vary in age, and their teaching experience ranges from 2 years to more than 30. They all have Master's degrees or higher, and all have a graduate specialization in the language in question. They are all professional language educators, even though a few are pursuing further graduate studies on a part-time basis. Some are foreign nationals, but all have experience in the U.S. educational setting. Almost all have had Oral Proficiency Interview familiarization training, and several have completed ACTFL OPI training. Many of them serve as Simulated Oral Proficiency Interview (SOPI) raters on a regular basis as well, and many have rated oral placement tests at the institution for a number of years. They have all attended training each year on how to administer and rate the tests. The amount of training per instructor thus varies with the number of years of experience in the progam. They all use the same set of rating scales, and all are of the norm-referenced categorical types, as adjusted for year 1 and year 2. Most of the instructor-raters have attended interrater reliability training sessions in which four or five actual student tapes have been rated by the group and in which expectations and interpretations of the criteria were discussed at length. Nevertheless, interaction in those sessions and other meetings of the group has highlighted on-going variation among the members in their expectations in the various categories being rated. The present survey should provide the opportunity to highlight specific areas of variation for further discussion and training.

The Instrument and Procedure

The author created a two-page checklist of 61 items relating to functions tested, formats used, and expectations held in rating (Appendix). There are 20 items on functions tested, thus sampling only a part of the curriculum in this area; 17 on formats used; and 24 on rater expectations of student performance. The instructors were asked to check all the statements that applied regarding their own procedures in testing and rating and their own expectations of learners in first- and second-year classes, which all of the instructors teach. In addition, they were provided space at the end of the questionnaire to write anything else they wished regarding their expectations of first- and second-year learners on speaking tests for their classes. The respondents were anonymous, since no place was provided on the survey for them to identify themselves. Anonymity was important, since one could assume that any in-

structors who felt their procedures or expectations did not match coordinated departmental expectations might not have wished to reveal them otherwise.

The instructors were accustomed to surveys and other efforts to research program functioning; therefore, they were simply asked via memo to fill out an attached survey in order to inform the coordinator's efforts to plan tester training for the semester in question. Forms were returned anonymously to the researcher's box over a period of days. Ninety percent of the instructor pool responded.

RESULTS AND DISCUSSION

As stated previously, respondents were asked to react to items covering functions tested, formats used, and expectations held. The results are shown in Table 1 and discussed in the following sections.

Functions

There was somewhat less variation among the instructor tester-raters in the area of functions tested than in the areas of formats and expectations. This finding might not seem surprising at first glance, given the coordinated nature of the program and the standard test samples distributed to faculty and learners; however, the same uniformity could have held true for formats but did not to the same extent.

The description function, which is a staple for Intermediate-level students and a logical starting point for Novices as well, was almost universally tested. All the instructors indicated that they included description of pictures and people, and 90% included description of places. Only 70% included description of objects, however.

Ninety percent of the instructor-raters had the learners ask questions on specific topics and get information about costs, times, and so on in real-life situations. Only 60% had students ask information questions about pictures. Interestingly, only 60% indicated that they had learners give information to others, although one would assume that participating in these question-asking formats would also include answering questions. Again, asking and answering information questions would seem appropriate functions for teaching and testing first- and second-year learners.

Seventy percent had learners roleplay greetings and introductions, and 70% had students express likes and dislikes. Both of these level-appropriate areas are part of the curriculum and of sample tests. Nevertheless, 30% of the instructors did not test them. In addition, 70% of the instructors had learners make requests as part of their roleplay; 30% did not, even though this possibility is also included in the sample tests.

All instructor-raters included narration in present and past tenses, and 90% included narration in the future tense. Again, the discrepancy is interesting, though small, as future-tense narration (be it formal or informal) is included in the sample tests. While present-tense narration could be considered appropriate for the learners' level of proficiency, past- and future-tense narration as Advanced-level tasks are in the realm of practice and goals more than of achievement and mastery.

Ninety percent of the instructors have the learners give directions for going somewhere and instructions for doing something, both of which are included in the sample tests but which vary in level appropriateness. Giving directions for going somewhere is considered Intermediate level, but giving instructions on how to do something can exceed the learners' level of proficiency, depending on the task or topic.

Only 60% of the instructor-raters include the comparison function on their tests, despite its inclusion in the curriculum and the sample tests. Only 60% included hypothesis, and 50% included persuasion. Forty percent included formal situations (work- or profession-related, for example), again despite the fact that there are such items available to them and such material is covered in both the first and second year. These functions are of the Advanced and Superior levels. Evidently some of these instructors have answered the question of how to rate learners on tasks that have been covered, but are beyond their level of proficiency, by eliminating the problem altogether and not including those functions on the speaking test at all.

Formats

There was somewhat greater variation among the instructor-raters on the questions regarding formats. Seventy percent used interviews, 50% used situations without complications, and 40% used situations with complications. Seventy percent used topics. Seventy percent used prepared material, referring to the sample tests distributed to students. Fifty percent also required the use of extemporaneous topics. Eighty percent expected performance at the phrase and sentence level; only 40% expected students to attempt performance at the paragraph level. Only 30% varied formats so that students would have to adapt their language to different registers. Seventy percent used formats that called for giving personal answers, not just general information, and 70% used formats that elicited variable answers. Given that as many as 30% of the instructors do not use some of the formats at all, one could assume that some learners are receiving more well-rounded assessment than others, if not also more well-rounded preparation.

The roleplay formats could be considered more difficult to perform (and to rate) than interviews, which consist of asking and answering questions, but roleplay can be an Intermediate-level format. Therefore, the fact that only half the instructors use roleplay is a concern because that format is the best simulation of real-life use of language. Having the learners speak extemporaneously is also more difficult than adhering to a specific repertoire of material; yet, it too is an essential skill that half of these learners are not attempting on tests. If only 40% of the instructor-raters expect students to attempt paragraph-level speech in the first 2 years of language study, this is yet another decision that has been made regarding what is too difficult for learners. The issue of how to rate the learners on beyond-level tasks does not arise for half of them because the

Table 1
Table of Responses
[Percentages of Positive Responses to Items by Instructor-Raters]

QUESTION THEMES	%	QUESTION THEMES	%	QUESTION THEMES	%
FUNCTIONS TESTED				Require students to speak without hesitation	30
Descriptions:		Situations without complications	50		
Pictures	100			Require students to use vocabulary covered	90
People	100	Situations with complications	40		
Places	90			Require students to use accurately all grammar covered	60
Objects	70	Topics	70		
Ask questions				Require accurate use of past, present, and future	80
Situations	90	Prepared material	70		
Pictures	60			Require students to handle any topics covered	80
		Extemporaneous material	50		
Answer Questions	60			Expect coherence	80
		Sentence -level formats	80		
Roleplay				Expect cohesion	50
Greetings	70				
Introductions	70	Paragraph-level formats	40	Expect sociolinguistic appropriateness	30
Express likes/ dislikes	70				
		Formats vary registers	30	Expect students to perform only Novice-Intermediate tasks well	50
Make requests	70				
		Require personal information	70		
Narration				Expect students to perform Advanced tasks well if covered	30
Present	100				
Past	100	Require variable answers	70		
Future	90				
		EXPECTATIONS		Expect students to perform Superior tasks well if covered	50
Giving directions	90				
		More than two errors allowed for an A grade	80		
Giving instructions	90				
		Require students to pronounce accurately	40		
Comparison	90				
		Require students to pronounce understandably	50		
Hypothesis	60				
Persuasion	50				
Formal situations, work-related	50				
FORMATS					
Interviews	70				

tasks simply are not required of them. Adapting language for different registers is also a beyond-level task, but one that is covered in the program. It appears in few of the speaking tests, apparently because it is also thought to be too difficult for the learners.

Instructor-Rater Expectations

The responses to the questions regarding instructor-raters' expectations of learners when rating revealed the greatest variation of all the variables. Eighty percent of the instructor-raters agreed that an A student could have more than one or two errors in a test speech sample, so they began on a similar footing. They were divided on pronunciation, however, with 40% indicating that learners must pronounce accurately and 50% indicating that learners should pronounce understandably but not necessarily entirely accurately. Thirty percent expected learners to speak without hesitation, which could be difficult for Novices even with the sort of semi-prepared repertoire testing used. Ninety percent expected learners to know and use the appropriate vocabulary that had been covered, and 60% expected the accurate use of all grammatical structures covered in the current semester and previously. This latter expectation is especially interesting, given that, as noted previously, many of the structures covered would not be mastered until the learners rose one or two more levels in their proficiency. In previous sections we saw that instructors omitted some functions and formats deemed too difficult; this type of exception occurs at about the same rate for grammar, which is nearly half the time. The one difference, however, at least for the grammar topics included, was the tenses, since accurate performance with past, present, and future was expected by 80% of the instructors.

Eighty percent of the instructors felt that learners should be able to handle all the topics covered. Eighty percent said they expected coherence, and 50%, cohesion, which are Advanced-level expectations. Thirty percent expected sociolinguistic appropriateness, which, while a low figure, nevertheless reflects a group of instructors who have another Advanced-level expectation for Novice and Intermediate learners. Seventy percent say they hold these expectations for the semi-prepared repertoire material, but only 10% hold these expectations for extemporaneous material, which may render the expectations somewhat more reasonable.

Half the respondents expected learners to perform well only on Novice and Intermediate material, and 30% expected them to perform well on Advanced material that had been covered. Only 10% expected learners to perform well on Superior material that had been covered. These responses are not entirely consistent with the percentage of instructors who expected Advanced-level grammar and functions, which was 60%. Thus, many instructors may expect higher-level functioning of students, even though only 30% of them at most marked these Advanced and Superior-level items.

Half the respondents agreed that Novices would perform fairly accurately because they were using

primarily memorized material and that Intermediates would perform relatively less accurately because they were now creating in the language. Apparently the other half of the respondents were not aware that the literature does appear to support those positions. Sixty percent agreed that students would make more errors on new material than on old material, a truism for which we would have expected a greater level of support. Sixty percent agreed that students would make more errors on extemporaneous material than on prepared material, where again we would have expected a higher level of support.

In the comments sections, the only respondents who provided additional information did not add categories of expectations. They merely reinforced the answers they had marked previously by elaborating on the reasons they expected students to speak without hesitation or the reasons they expected grammatical accuracy.

CONCLUSIONS

Clearly, this study reinforces findings in the literature that even with seemingly well-defined expectations for students in syllabi and clarification of expectations through discussion and training for faculty, testing and rating procedures can and do vary. The entire area of beyond-level material evidently needs to be discussed more carefully in this program and most likely in any university or high-school program in which the issue has not yet been raised. Those areas that are being covered in the courses need to be tested, and clarification is needed about how to include them and rate them appropriately. Language-specific expectations for pronunciation and grammar need to be discussed in some detail, just as Richards and Chambers (1996) found. In particular, expectations regarding the use of past and future tenses require further definition, and expectations regarding other structures need to be explored. Instructions, comparisons, hypothesis, and persuasion are other areas for which discussion is indicated. The concerns about fairness to learners that instructors may be expressing by omitting some areas from tests should be addressed so that adaptation and not elimination is the solution. Clarifying and modifying rating rubrics is essential to ensuring greater agreement on what instructors are expecting.

In addition, there clearly needs to be a broader use of varied formats in the testing so that learners form a broader communicative base and so they are not affected by always having to perform in their weakest format, should that be the case. Instructors should include formats that call for adaptation to the situation and interlocutor. As noted, roleplay is often neglected, and it is the format that best affords opportunity to negotiate meaning and attend to sociolinguistic details (Omaggio, 1980). Teachers should give learners opportunities to perform at the paragraph level so that they can be led in that direction. They should ask students to perform extemporaneously as well as with their prepared repertoire in order to facilitate learners' ability to use the language in the real world.

Perhaps we need to develop brief tester-rater manuals for use within our programs with content based on faculty consensus from discussion, training, and further study of the literature. In addition, as Kenyon and Stansfield (1993) suggest, the creation of a set of reference tapes for use within a program could be very beneficial. If instructors had some sample student responses and ratings for the different formats used on each of the tests in each of the courses in the program, their integration into the process when newly hired and their on-going comfort and consistency would be better ensured.

Further discussion and training, including inter-rater reliability training, is needed on all these topics. Even though there will undoubtedly always be some degree of variation from one rater to another, we as professionals have focused heavily on proficiency testing and rather little on prochievement or classroom testing. With an increased focus on the quality of our ratings in the tests we give most frequently, we can only enhance the effectiveness of our programs and our students' achievement and proficiency.

REFERENCES

Byrnes, H. (1987). Proficiency as a framework for second language acquisition. *The Modern Language Journal, 71* (1), 44-49.

Chambers, F., & Richards, B. (1995). The free conversation and the assessment of oral proficiency. *Language Learning Journal, 11*, 6-10.

Clark, J. (1989). Multipurpose language tests: Is a conceptual and operational synthesis possible? *Language teaching, testing, and technology*. Washington, D. C.: Georgetown University Press.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language International*, 125-143.

Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing, 13* (1), 23-52.

Gonzalez Pino, B. (1989). Prochievement testing of speaking. *Foreign Language Annals, 22* (3), 478-487.

Hart Gonzalez, L. (1994). Raters and scales in oral proficiency testing: The FSI experience. Paper presented at the Sixteenth Annual Language Testing Research Colloquium, Washington, D.C.

Kenyon, D., & Stansfield, C. (1993) Evaluating the efficacy of rater self-training. *Language Testing Research*. Cambridge: The Fifteenth Annual Language Testing Research Colloquium.

Meredith, R. (1990). The oral proficiency interview in real life: Sharpening the scale. *Modern Language Journal, 74* (3), 288-296.

Mullen, K. (1978). Direct evaluation of second language proficiency: The effect of rater and scale in oral interviews. *Language Learning, 28* (2), 302-308.

Omaggio, Alice. (1980). Priorities for the 1980s. In D. Lange (Ed.), *Proceedings of the National Conference on Professional Priorities* (pp. 47-53). Hastings-on-Hudson, NY: ACTFL.

Richards, B., & Chambers, F. (1996). Reliability and validity in the GCSC oral examination. *Language Learning Journal, 14*, 28-34.

Ross, S. (1987). An experiment with a narrative discourse test. *Language Testing Research*. Monterey, CA: The Ninth Annual Language Testing Research Colloquium. Monterey, California.

Stansfield, C., & Kenyon, D. (1992). The development and validation of a Simulated Oral Proficiency Interview. *Modern Language Journal, 76* (2), 129-141.

Thompson, I. (1995). A study of interrater reliability of the ACTFL Oral Proficiency Interview in five European languages: ESL, French, German, Russian, and Spanish. *Foreign Language Annals, 28* (3), 407-422.

Thompson, I. (1996). Assessing foreign language skills: Data from Russian. *Modern Language Journal, 80* (1), 47-65.

Young, R. (1992). *Expert-novice differences in oral foreign language proficiency*. Paper presented at the Colloquium on Non-Native Speaker International Discourse at the Fourteenth Annual Meeting of the American Association for Applied Linguistics, Seattle, Washington.

APPENDIX
ADMINISTERING AND RATING ORAL TESTS

Check all that apply in your point of view and in the way that you administer and rate oral tests.

1. In oral tests (tests of speaking) for my first- and/or second-year foreign language students, at some point during the year the students must

   _ describe pictures
   _ describe objects
   _ describe people
   _ describe places
   _ ask questions based on a picture
   _ ask questions on topics, such as family, studies, etc.
   _ ask questions to get information about cost, times, etc.
   _ express likes and dislikes
   _ greet others, perform introductions, say farewell
   _ make requests
   _ give information on a varietv of topics
   _ narrate in the present; e.g., say what they do on weekends, during a typical day, etc.
   _ narrate in the past; e.g., say what they did on a weekend, holiday, typical day, etc.
   _ narrate in the future; e.g., say what they will do on a holiday, in their

future work, etc.
- _ explain a process, such as how to make a particular dish
- _ give directions or instructions., such as how to go from one place to another
- _ compare two (or more) pictures, people, places, objects
- _ say what they would do in a hypothetical situation
- _ try to persuade someone of something
- _ use formal language (introduce a speaker, start a formal talk, explain an abstract topic, such as socialism, etc.)

2. When taking speaking tests, my first- and/or second-year foreign language students are expected to

- _ interview a partner (another student or the teacher)
- _ roleplay situations without complications
- _ roleplay situations with complications
- _ perform extemporaneously sometimes
- _ perform with prepared situations, topics, presentations, etc.
- _ vary the way they speak to suit the audience (listener) and the situation (be more or less formal)
- _ speak at the phrase level
- _ speak at the sentence level
- _ give personal answers based on their own information, experiences, preferences, opinions, etc.
- _ give variable answers (answering open-ended questions rather than those which have only one right answer, such as what day it is)
- _ answer closed questions (those with one right answer)
- _ speak in context (on a particular topic or situation)
- _ handle a random selection of topics, questions, situations, etc., from a pool of them which have been practiced and/or prepared during the testing period

3. When I rate the speaking tests of my foreign language students in first or second-year courses, I expect students to perform as follows for a grade of A:

- _ have only one or two errors in the speech sample
- _ pronounce accurately
- _ pronounce understandably, but not always accurately
- _ speak virtually without hesitation
- _ know and use the appropriate vocabulary (with no English)
- _ use accurately the grammatical structures that have been covered in this level class and in previous levels
- _ complete adequately all the types of tasks or functions that have been covered
- _ talk adequately about any and all of the content areas that have been

covered (family, clothing, current events, jobs, etc.)
- _ speak in a culturally or sociolinguistically appropriate manner
- _ organize their thoughts logically
- _ transition appropriately from one idea to another
- _ do all of the above when speaking with prepared material
- _ do all of the above when speaking extemporaneously
- _ perform well only on Novice tasks (ACTFL scale)
- _ perform well only on Intermediate tasks
- _ perform well on Advanced tasks if they have been covered
- _ perform well on Superior tasks if they have been covered
- _ accurate performance from Novices because they are memorizing their material
- _ less accurate performance from Intermediates because they are now creating in the language
- _ more errors on new material than on old
- _ more errors on extemporaneous formats
- _ fewer errors on prepared formats
- _ phrase- or sentence-level or length performance or responses
- _ paragraph-level or length performance or responses

4. Other: Your additional comments about what you expect from first and/or second-year foreign language students on their speaking tests: